E. T. JAYNES

SOME RANDOM OBSERVATIONS

INTRODUCTION

When told of the plan to devote an issue of *Synthese* to the principle of maximum entropy (PME), my first reaction was puzzlement as to why philosophers should be so interested in it. Then a plausible reason appeared to be that PME has been perceived (correctly, in my view) as a small but nonnegligible part of a fundamental revolution in thought now taking place.

We have the analysis of Kuhn describing how new foundation concepts make their way into science; but today in order to study this one does not need to go back to the history books to read of epicycles and ellipses. We can observe in our midst the phenomenon of one paradigm in the process of being replaced by another.

In one respect, our present revolution is not like those of cosmology, evolution, or relativity – which, however grand their concepts, were specialized to one particular area of science. What we lack in grandness of concept we make up for in generality; the new revolution concerns the principles of all human inference, and it applies with equal force to all areas of science. The conceptual disorientation and resulting controversy are being exhibited now in the scientific journals of half a dozen different fields.

But we have to admit that the spectacle unfolding today teaches us very little that could not have been learned from those history books; it is astonishingly like what happened in the time of Galileo, down to such small details as to give one a spooky feeling. However, this is not the place to develop that theme.

HOW THE REVOLUTION STARTED

It should be pointed out that my publications, starting in 1957, make a rather late entry into this movement. Prior to that, many important works – by B. de Finetti, H. Jeffreys, R. T. Cox, C.E. Shannon, I. J. Good, A. Wald, R. Carnap, L. J. Savage, and D. V. Lindley – had

Synthese **63** (1985) 115–138. 0039–7857/85.10
© 1985 *by D. Reidel Publishing Company*

started this general shift in thinking. These works had laid the basis of the revolution by their demonstrations that probability theory can deal, consistently and usefully, with much more than "frequencies in a random experiment".

In particular, Cox (1946) proved by theorem what Jeffreys (1939) had demonstrated so abundantly by example; the equations of probability theory are not merely rules for calculating frequencies. They are also rules for conducting inference, uniquely determined by some elementary requirements of consistency. de Finetti, Wald, and Savage were led to the same conclusion from entirely different viewpoints.

As a result, probability theory, released from the frequentist mold in which Venn, von Mises, and Fisher had sought to confine it, was suddenly applicable to a vast collection of new problems of inference, far beyond those allowed by "orthodox" teaching. One might think that an explosive growth in new applications would result. But it did not, until quite recently, because half of probability theory was missing.

Orthodox statistics had developed only means for dealing with sampling distributions and did not even acknowledge the existence of prior probabilities. So it had left half of probability theory – how to convert prior information into prior probability assignments – undeveloped. Jeffreys (1948) recognized that this half was a necessary part of any full theory of inference, and made a start on developing these techniques. Indeed, his invariance principle was closely related to PME, starting from nearly the same mathematical expression.

Much earlier, both Boltzmann (1877) and Gibbs (1902) had invoked the mathematical equivalent of PME as the criterion determining Boltzmann's "most probable distribution" and Gibbs' "grand canonical ensemble". But this fact had been completely obscured by generations of textbook writers who sought to put frequentist interpretations on those results, not realizing that in so doing they were cutting Statistical Mechanics off from generalizations to nonequilibrium problems.

It was only at this point that I entered the field, with the suggestion that Shannon's Information Theory provided the missing rationale for the variational principle by which Gibbs had constructed his ensembles. In my view, Gibbs was not assuming dynamical properties of an "ergodic" or "stochastic" nature, but only trying to construct the "most honest" representation of our state of information. One can find much support for this view in Gibbs' own words.

Pragmatically, in equilibrium problems this could not lead to any new

results because PME yielded the same algorithm that Gibbs had given and which was already in use as the basis of our calculations. The value of the principle lay rather in the realization of its generality; when seen in this light it was clear that the Gibbs "canonical ensemble" formalism was not dependent for its validity on the laws of mechanics; it would apply as well to any problems of inference, in or out of physics, that fit into the same logical format.

The place that PME occupies in our present revolution is that it is one of the principles that has proved useful – and fairly general – in the task of developing that missing half (logically, the first half) of probability theory. Other such principles are group invariance, marginalization, coding theory, and doubtless others not yet thought of; this appears to be a fertile area for research.

Turning to the present discussions of Bayesian methods and PME, there is an appalling gulf between what scientists are doing with them and what philosophers are doing to them. What is needed most is not still more contention over the exact meaning of things that were written thirty years ago, but a kind of report from the outside world on what has happened since, how these methods have evolved, and what they are accomplishing today. However, since I was asked to comment on other works published in this issue and elsewhere, only a little of this can be given here.

It is of course distressing that a few philosophers disapprove of the work of so many scientists, engineers, economists, and statisticians on Bayesian/PME methods, out of misunderstanding. However, I can speak only for myself, and even then, if thirty commentators interpret one's work in thirty different ways it is not feasible to analyze them all and write – or read – thirty separate replies. Therefore, I can only restate my position and some technical points as clearly as possible, then reply to a few specific comments where it appears that failure to do so would encourage further confusion.

TERMINOLOGY

One critic states that my terminology is nonstandard and can mislead. He fails to note that this applies only to the 1957 papers; and even there my terminology was standard when written. It is, for example, essentially the same as that used by Jimmie Savage (1954). It is not our fault that Latter-Day Commentators, in ill-advised attempts to "classify

Bayesians", have scrambled the meanings of our words, producing a language appropriately called NEWSPEAK. To translate, we note a few approximate equivalences between standard terminology of the 1950's and NEWSPEAK:

1950'S	NEWSPEAK
objective	orthodox
subjective	objective
personalistic	subjective
Bayes' theorem	conditioning
conditioning	ancillarity

To this we may add the alarming spread in use of the terms "prior distribution" and "posterior distribution" (which had clearly established meanings as referring to the application of Bayes' theorem) to describe instead two different maximum entropy distributions. As a result, some writers are now unable to distinguish between PME and Bayes' theorem and are led thereby into nonsensical calculations and claims.

PME belongs to that neglected first half of probability theory; Bayes' theorem to the second half that starts only after a prior has been assigned by PME or one of the other principles (as we shall see, the second half, given one prior, can then determine other priors consistent with it).

Because of this utter confusion that has been visited upon us, it is today misleading to use the terms "subjective" and "objective" at all unless you supply the dictionary to go with them. My more recent works have used them only in such phrases as "subjective in the sense that —."

My papers of 1957 used the term "subjective" not only in the superficial sense of "not based on frequencies", but in a deeper sense as is illustrated by the following statement of position:

In relativity theory we learn that there is no such thing as "absolute" or "objective" simultaneity. Nevertheless, each observer still has his "subjective" simultaneity, depending on his state of motion; and this, being a consequence of using a coordinate system, is just as necessary in describing his experiences as it was before relativity. The advance made by relativity theory did not, then, lie in rejecting subjective things; but rather in recognizing that subjective character, so that one could make

proper allowance for it, the "allowance" being the Lorentz transformation law, which shows how the coordinates we assign to an event change as we change our state of motion.

It has seemed to me from the start, that a scientist should have just the same attitude toward probability. There is no such thing as "absolute" or "physical" probability; that is just as much an illusion as was absolute simultaneity. Yet each of us still has his "subjective" probability, depending on (or rather, describing) his state of knowledge; and this is necessary for conducting his reasoning. Achievement of rational thinking does not lie in rejecting "subjective" probabilities, but rather in recognizing their subjective character, so that we can make proper allowance for it, the "allowance" being Bayes' theorem, which shows how the probability we assign to an event changes as we change our state of knowledge.

The phrase "reasonable degree of belief" originates from Jeffreys, although some try to connect it to me. My admiration for Jeffreys has been expressed sufficiently (Jaynes, 1980); but his terminology created some of his difficulties. That innocent looking little word "reasonable" is to some readers as the red flag to the bull; the adrenalin rises and they miss the substantive content of his message. Jimmie Savage (1954) used instead "measure of trust" which I felt had both economic and emotional overtones. Therefore I have used the less loaded phrase "degree of plausibility" which seems to express more accurately what inference really involves.

PROBABILITY AND FREQUENCY

One commentary notes, with commendable perceptiveness, that I do not "disallow the possibility" of the frequency interpretation. Indeed, since that interpretation exists, it would be rather hard for anyone to deny the possibility of it. I do, however, deny the necessity of it; and this is a change from the position stated in my 1957 papers. At that time, I was under the impression that some applications required a frequency interpretation, simply because my professors and textbooks had said so, and had not yet fully shaken off their authority. It required a few more years of experience to realize that every connection between probability and frequency that is actually used in applications, has proved on analysis to be derivable as a consequence of the Laplace–Jeffreys–Cox form of probability theory.

In my terminology, a *probability* is something that we *assign*, in order to represent a state of knowledge, or that we *calculate* from previously assigned probabilities according to the rules of probability theory. A *frequency* is a factual property of the real world that we *measure* or *estimate*. In this terminology, the phrase "estimating a probability" is just as much a logical incongruity as "assigning a frequency".

The fundamental, inescapable distinction between probability and frequency lies in the aforementioned relativity principle. Probabilities change when we change our state of knowledge; frequencies do not. It follows that the probability p that we assign to an event E can be equal to its frequency f only for certain particular states of knowledge. Intuitively, one would expect this to be the case when the only information we have about E consists of its observed frequency; and the mathematical rules of probability theory confirm this in the following way.

We note the two most familiar connections between probability and frequency. Under the assumption of exchangeability and certain other prior information (Jaynes, 1968, or Skilling's presentation in this issue), the rule for translating an observed frequency in a binary experiment into an assigned probability in Laplace's rule of succession. Under the assumption of independence, the rule for translating an assigned probability into an estimated frequency is Bernoulli's weak law of large numbers (or, to get an error estimate, the de Moivre–Laplace limit theorem).

However, many other connections exist. They are contained, for example, in the principle of transformation groups (Jaynes, 1971), in the PME formalism, and in the theory of random fluctuations (Jaynes, 1978).

If anyone wished to research this matter, I think he could find a dozen logically distinct connections between probability and frequency, that have appeared in various applications of PME, transformation groups, and Bayes' theorem. But these connections always appear automatically, whenever they are relevant; there is never any need to *define* a probability as a frequency.

Indeed, Bayesian theory may justifiably claim to use the notion of frequency more effectively than does the "frequency" theory. For the latter admits only one kind of connection between probability and frequency, and has trouble in cases where a different connection is appropriate. Those cases include some important, real problems which

are today at the forefront of new applications.

Today, Bayesian practice has far outrun the original class of problems where frequency definitions were usable; yet it includes as special cases all the useful results that had been found in the frequency theory. In discarding frequency definitions, then, we have not lost "objectivity"; rather, we have advanced to the flexibility of a far deeper kind of objectivity than that envisaged by Venn and von Mises. This flexibility is necessary for scientific inference; for most real problems arise out of incomplete information, and have nothing to do with random experiments.

BUT WHAT ABOUT QUANTUM THEORY?

Those who cling to a belief in the existence of "physical probabilities" often react to the above arguments by pointing to quantum theory, in which physical probabilities appear to express the most fundamental laws of physics. Indeed, it was just this circumstance that first aroused my interest in the philosophy of probability theory. But it needs to be emphasized that present quantum theory uses entirely different standards of logic than does the rest of science.

In biology or medicine, if we note that an effect E (for example, muscle contraction, phototropism, digestion of protein) does not occur unless a condition C (nerve impulse, light, pepsin) is present, it seems natural to infer that C is a necessary causative agent for E. Most of what is known in all fields of science has resulted from following up this kind of reasoning. But suppose that condition C does not always lead to effect E; what further inferences should a scientist draw? At this point the reasoning formats of biology and modern physics diverge sharply.

In the biological sciences one takes it for granted that in addition to C there must be some other causative factor F, not yet identified. One searches for it, tracking down the assumed cause by a process of elimination of possibilities that is sometimes extremely tedious. But persistence pays off; over and over again medically important and intellectually impressive success has been achieved, the conjectured unknown causative factor being finally identified as a definite chemical compound. The acetyl-choline, auxin, and trypsin were found in this way. Most enzymes, vitamins, viruses, and other biochemically active substances owe their discovery to this reasoning process.

In modern physics, we do not reason in this way. Consider, for

example, the photoelectric effect. The experimental fact is that the electrons do not appear unless light is present. So light must be a causative factor. But light does not always produce ejected electrons; even though the light from a unimode laser is present continuously, the electrons appear only at particular times that are not determined by any known parameters of the light. Why then do we not draw the obvious inference, that in addition to the light there must be a second causative factor, still unidentified, and the physicist's job is to search for it?

What we do today is just the opposite; when no cause is apparent we postulate that no cause exists – ergo, the laws of physics are indeterministic and can be expressed only in probability form. There is no "auxin" for electrons: the light determines, not whether a photoelectron will appear, but only the probability that it will appear.

Biologists have a mechanistic picture of the world because, being trained to believe in causes, they continue to search for them and find them. Physicists have only probability laws because for two generations we have been trained not to believe in causes – and so we have stopped looking for them. To explain the indeterminacy in current physical theories we need not suppose there is any indeterminacy in Nature; the mental attitude of physicists is already sufficient to account for it.

I suggest, then, that those who try to justify the concept of "physical probability" by pointing to quantum theory, are entrapped in circular reasoning. Probabilities in present quantum theory express the incompleteness of human knowledge just as truly as did those in classical statistical mechanics; only its origin is different.

In classical statistical mechanics, probability distributions represented our ignorance of the true microscopic coordinates – ignorance that was avoidable in principle but unavoidable in practice, but which did not prevent us from predicting reproducible phenomena.

In the Copenhagen interpretation of quantum theory, probabilities express the ignorance due to our failure to search for the real causes of physical phenomena. This may be unavoidable in practice, but in our present state of knowledge we do not know whether it is unavoidable in principle; the Copenhagen "central dogma" simply asserts this, and draws the conclusion that belief in causes, and searching for them, is philosophically naive.

The deepest driving motivation behind all my work on statistical theory is not just the desire for more powerful practical methods of inference. It is rather the conviction that progress in basic understand-

ing of physical law, prevented for fifty years now by the positivist Copenhagen philosophy, can be resumed only by a drastic modification of the view of the world now taught to physicists.

Present quantum theory contains an enormous amount of very fundamental truth; but its mathematics describes in part physical law, in part the process of human inference, all scrambled together in such a way that nobody has seen how to separate them. Many years ago, I became convinced that this unscrambling would require that probability theory itself be reformulated so that it recognizes explicitly the role of human information and thus restores the distinction between reality and our knowledge of reality, that has been lost in present quantum theory. Bayesian probability theory appears to be the only form capable of doing this.

TIME CHANGES ALL THINGS

One of the complicating factors in this discussion is that commentaries are addressed to work done over many years. Of course, my own views about probability theory have changed in this time. As many readers know, I have been trying to finish a book on probability theory since 1956, but cannot because new, exciting things are still being discovered too fast to write up. In fact, a new caution has been forced on my current writing, because my views on probability theory are today in process of more rapid change than at any time in the past.

Several recent events have conspired to bring this about. New applications of maximum entropy, new facts unearthed in marginalization theory, discovery of the combinatorial work of Gian-Carlo Rota, some belated acquaintance with computer design and programming, that changed my perception of the realities of implementation – all have displaced my thinking from its settled position of ten years ago.

Another factor has been my growing dismay at the way young people are, like lemmings, rushing headlong into a sticky swamp of paradoxes (nonconglomerability, Borel-Kolmogorov, marginalization, improper priors, finite vs. countable additivity, etc.) and their seeming inability to recognize that these have nothing to do with real substantive issues: In most cases, they demonstrate only the need for greater caution in approaching infinite sets.

The projected book was held up for a long time because the theory seemed incomplete; a major aim was to settle this 150-year-old

controversy by refinements of the theorems of R. T. Cox (1961), showing that the product and sum rules of probability theory are the only consistent rules for conducting inference. In this endeavor, everything could be established in a satisfactory way for probabilities on finite sets, but I could not find a neat, elegant argument that would do this as well for infinite and continuous sets.

After years of failure, and then seeing Rota consistently doing things right, and the lemmings consistently doing things wrong, the suspicion grows that the generalization I have been seeking may not, after all, exist. It is beginning to appear as if all correct results in probability theory are either combinatorial theorems on finite sets or well-behaved limits of them.

Evidence for this was accumulated very slowly and painfully, but eventually there came the insight that makes it possible to summarize it all in one paragraph. There is a single technique by which all the aforementioned paradoxes – and any number of new ones – can be manufactured. It requires only three steps: (a) Start from a correct, well-behaved result on a finite set; (b) Pass to the limit of an infinite set without specifying how the limit is approached; (c) Ask a question whose answer depends on how the limit was approached.

This procedure is guaranteed to produce a paradox, in which a seemingly reasonable question has more than one seemingly right answer. For example, nonconglomerability is an artifact created thus: (a) Start from a finite two-dimensional $(N \times M)$ array of probabilities $P_{ij}(1 \leq i \leq N;\ 1 \leq j \leq M)$; (b) Pass to the limit $N \to \infty$, $M \to \infty$ without specifying the limiting ratio (N/M). (c) Ask for the probability, on the infinite set, of the event $(i > j)$. If one looks only at the limit, and not the limiting process, the source of the difficulty is concealed from view and may not be discovered for a long time.

More generally, these paradoxes are simply the result of trying to jump directly into an infinite set without considering any limiting process from a finite set. The belief that infinite sets possess some kind of "existence" and mathematical properties in their own right, independently of the limiting process that defines them, is very much like the belief in absolute probability or absolute simultaneity. Cantor and Hilbert apparently held some such belief. But Gauss, Kronecker, Poincaré, Brouwer, and Weyl have all warned against this kind of reasoning; perhaps it is time to heed them.

On the other hand, the consistency theorems of Cox apply to finite

sets and nobody has been able to produce any paradox as long as he applies the product and sum rules of probability theory, derived by Cox, on finite sets. It thus appears that the finite discrete theory of inference that I had developed in 1956, but thought to be incomplete, may be after all the whole thing. It seemed inelegant to define the entropy of a continuous distribution from the limit of a discrete one; but perhaps that is the only justification that exists. And aesthetically, with the development of a little bit of computer mentality, discreteness no longer seems as ugly as it once did.

Of course, these remarks are offered only as interesting (or perhaps provocative) conjectures; and not as revelations of a startling new discovery about the nature of probability theory. But it seemed advisable to mention this evolution of thinking, so that others may better judge whether to dwell on views expressed long ago.

INTUITIVE PARADOXING

Another of the popular games that people play with Bayesian inference is a variant of the lemming procedure. Instead of using ill-defined infinite sets to generate a paradox between two mathematical results, one may equally well manufacture a paradox between the mathematical theory and his own intuitive commonsense judgments. The vehicle for this is the improper (i.e., nonnormalizable) prior probability distribution – again, considered as an object in its own right without regard to the limiting operation that defines it.

Passage from a probability distribution on a finite set to a proper continuous distribution is almost always uneventful, leading to no paradox but rather to some of the most important results of probability theory. But passage from a proper continuous distribution to an improper one (such as a uniform density $p(f)$ extending all the way to $\pm\infty$, or a Jeffreys prior density $1/f$ extending to $f=0$), may lead to useful results or to paradoxes, depending on further details of the problem, including the particular data set D that has been observed.

Study of these matters has convinced me that the Cox finite set approach gives us a theory of probability that is general enough for all real problems and free of paradoxes (in contrast, to define probability in terms of additive measure leaves one at the mercy of all the paradoxes of infinite set theory, which are irrelevant for real problems). In this theory, rules of calculation with improper distributions are not defined

except as a limiting form from a sequence of proper distributions – and then only when that limit proves to be mathematically well-behaved.

Again, the paradoxes can be manufactured if one violates these rules by jumping directly into an improper prior without bothering to check whether it is a well-behaved limit from a proper one.

It is a common experience, in the present primitive state of development of that neglected half of probability theory, that some proposed prior distribution looks harmless, until analysis shows that it has unexpected consequences. Typically, one finds that it leads to a posterior distribution that our intuitive common sense rejects quite forcefully. Supposing the sampling distribution and data established beyond question, this leaves two possibilities: (A) intuition is faulty, and needs to be re-educated; (B) intuition is sound, but it is using prior information different from that expressed by the prior probability distribution.

As Skilling notes, the $1/f$ prior cannot be accepted as generally valid. It hardly requires numerical examples to show this, for if we know that $f > 0$ but we assign a prior distribution that diverges as $f \rightarrow 0$, we have only to perform an experiment (i.e., obtain a particular data set D) that does not exclude arbitrarily small values of f; and there is our paradox. For our inferences about f from such data then depend, necessarily, on our prior information concerning the possibility of very small values of f; but our prior distribution has thrown this away.

Indeed, it seems a platitude that if the data D are uninformative about any question Q (arbitrarily small f being one example), then our posterior opinions about Q will depend, necessarily, on our prior opinions about Q. If we assign a prior distribution that represents an absurdly impossible state of knowledge concerning Q, then that absurdity, if uncorrected by the data, must remain in the posterior distribution.

But it seems strange that, if one knows in advance what question Q he is going to ask, he would perceive the absurdity only after seeing data D that are uninformative about Q! Why did he not perceive it before seeing such data? This is a psychological peculiarity that we cannot account for at present.

Let us extend ths observation. Given any improper prior $p(f|I)$, one can find a question Q to which $p(f|I)$ gives an absurd answer. If the sampling distribution is sufficiently broad, one can also find a conceiv-

able data set D which is so uninformative about Q that the absurdity will remain in the posterior distribution $p(f|DI)$. So we have a rich field for the playing of this game; one can mass-produce as many paradoxes as one pleases.

Skilling has examined a few improper priors in a few circumstances, and duly finds examples of this phenomenon. From this evidence, he feels able to conclude: "... no prior can be found which is usable for all experiments ...."

From other evidence, I conjecture instead: "Any proper prior $p(f|I)$ represents a conceivable state of prior knowledge; given that and any experiment $p(D|fI)$, the posterior distribution $p(f|DI)$ will represent the reasonable conclusions from that prior information and data. Paradoxing ceases when impropriety ceases".

This does not mean that we should *never never* use an improper prior. Quite the contrary, most of the useful results of Bayesian inference have been obtained with improper priors. It means, rather, that before using an improper prior we need to check its suitability. This depends on the question we are asking and the sampling distribution (for example, there can be no objection to the uniform prior if we are estimating the mean of a normal distribution of known variance; for then the evidence of a single data point is enough to cancel the absurdity).

My conjecture is that a proper prior is respectable enough to go anywhere, and will not embarrass you whatever experiment you perform or whatever question you ask.

But the above remarks, while warning us of the bad roads, do not direct us to the good ones; can we now become more constructive and show how, in the present state of that neglected half of probability theory, one can actually find reasonable priors for the problems that Skilling describes?

We cannot go into specific technical details here, because that requires knowledge of the subject matter (astronomy and instrumentation) far beyond what I possess. At the time of writing (January 1983) it appears that John Skilling and I may soon be working together on these problems, and in another year or so we shall both be much better informed about them. But I can indicate here a general line of reasoning that starts us off on one of the good roads, aware that there may be others equally good.

Actually, Skilling and I are not nearly as far apart in our views as might appear superficially. But a point of principle can be expounded more clearly in a situation of greater contrast; and so I want to magnify that difference in the form of a Galilean Dialogue between two imaginary scientists.

Mr. A, unlike Skilling, has only orthodox statistical training and no Bayesian experience at all; but he does share Skilling's interest in estimating a parameter f of the Crab nebula. Mr. B is not an astronomer, but does share some of my experience in the analytical construction of priors.

A GALILEAN DIALOGUE

Prologue

Mr. A has been advised to use Bayesian methods, and as the scene opens he is expressing the standard doubts of nonBayesians.

A: "My prior information I obtained about the Crab nebula is just too vague. No definite prior probability distribution $p(f|I)$ can be based on it."

B: "Can you think of some additional information which, if you had it, would help to make the problem more definite?"

A: "Of course. If only I knew the total energy flux T of the Crab nebula, then the probability $p(f|TI)$ would be a determinate quantity."

B: "Excellent. You perceive a quantity T, not yet incorporated into the problem, that would be relevant. Then let's just introduce T into the conversation; using the sum rule, then the product rule, we have:

$$p(f|I) = \int p(fT|I)dT = \int p(f|TI)p(T|I)dT. \qquad (\dagger)$$

Now, then, what do you know about T?"

A: "Nothing. That's just why . . . ."

B: "But you must know something, else how could you know that T exists and is relevant? Surely you know it is bounded; after all, we aren't being scorched by the radiation from the Crab nebula.".

A: "Well, yes, but . . . ."

B: "And surely you know that it is not zero, else we would never be thinking about this problem."

A: "Well, yes — but I don't have any numbers, so how can that help?"

B: "But you do have some numbers, just think a bit harder. You know that $T < 10^{60}$ watts, because that would drain off the entire $E = mc^2$ energy of the Crab nebula in a microsecond. And you know that $T > 10^{-60}$ watts, because that is less than one microwave photon in the age of the universe, and nobody outside would know the Crab nebula exists. So you have two numbers, $T_{\min}$ and $T_{\max}$, within which you know the true value surely lies."

A: But this is silly. Such vague stuff can't help me!"

B: "*O ye of little faith*!" How do you know such stuff can't help until you give it a chance? Actually, knowing that $\log_{10} T$ is confined to a finite interval, $(-60, +60)$, is saying quite a lot. It makes the integral (†) converge, so now we can get a definite prior $p(f|I)$, which a moment ago you thought was impossible. We have made real progress."

A: "But the result is going to depend on those ridiculous numbers, 10^{60} and so on. They don't mean anything."

B: "Wait—we're not yet finished. How do you know the result depends on $T_{\min}$ and $T_{\max}$ until you study the matter? Let me ask a new question: Suppose you knew f; could you then estimate T?"

A: Of course. $p(f|TI)$ is a definite, calculable sampling distribution and as a function of T it is the likelihood function. I could get a maximum likelihood estimate of T, and find the accuracy of the estimate by setting up a 90% confidence interval. And I can do all this without ever mentioning those useless prior probabilities."

B: "Would that confidence interval extend out to $T_{\min}$ or $T_{\max}$?"

A: "Of course not. Those are, as I said, ridiculous numbers by many orders of magnitude."

B: "Then, in fact, the integral will converge so well that the result, if we round it off to four or five significant figures, won't depend on what those ridiculous numbers are."

A: "Well, yes, that's right. But I have no idea what $p(T|I)$ is in the region where it does contribute to the integral (†), so we'll still have no definite solution."

B: "But we're still not finished. You said you don't have any idea of what T is. Now how wide is that likelihood function going to be?"

A: (pause) "Actually, it's very broad. It would cover about a 2 : 1 range

so I couldn't estimate T very accurately from f, after all. So it's clear that your line of reasoning is not going to help us."

B: "Actually, 2:1 is quite a small range, only 0.3 on our logarithmic scale of (−60, +60). But surely the prior density $p(T|I)$ that describes your prior information about T does not vary widely in an interval of 0.3; if it did, that would imply some rather definite prior information about T, and you said a moment ago that you know nothing about its value."

A: "Yes, that's right – Oh – now I see the point."

B: "So now perhaps we are finished."

The Moral

To find prior probabilities we do, of course, need some actual prior information; merely proclaiming "complete ignorance" is not enough. But a surprisingly small amount of additional information suffices. If you can find any quantity T that is relevant to the inference (i.e., sufficiently relevant to make its likelihood integrable), then by the method (†) extremely vague information about T will give you a definite prior probability $p(f|I)$ for f; something that anti-Bayesians think is impossible to determine. Knowing the numerical value of T is not necessary; the mere qualitative fact of its existence and finiteness is already sufficient.

This is a very important consequence of probability theory, showing how with a little input $p(T|I)$ from that neglected first half, the Bayesian second half can, so to speak, take over the job of the first half and construct prior probabilities for other quantities. But it is something which statisticians could have learned from Harold Jeffreys in the 1930's, had they been willing to listen to him.

At this point Mr. A, slightly put off by the above turn of events, overhears this and returns to the attack; our dialogue resumes;

A: "But wait a minute, we're not finished. There isn't any definite solution because maybe there's another quantity U that is also relevant. If I used U instead of T, surely I wouldn't get the same result $p(f|I)$ in general."

B: "Of course you wouldn't; nor should you, for you would be solving a different problem. Knowing about T but not U, is a different state of knowledge from knowing about U but not T."

A: "But now your whole system crashes after all; for how can I ever know which quantities to take into account, or which other ones may be lurking out there unrecognized? There are no definite Bayesian inferences."

B: (wearily) "The tenacity and determination with which the orthodox mind resists simple common sense is truly amazing. I have now to point out a principle of morality, then one of pragmatism. *Morality*: In principle, one should always take into account all the prior information he has. To withold pertinent information from the theory, and then blame the theory for giving unsatisfactory results, is intellectual dishonesty. If you know about both T and U, then you should take them both into account by using their joint distribution in (†). *Pragmatism*: In the real world, the problem we face is, necessarily, to do the best we can with the information we have. If someone better informed than I takes into account further information of which I am unaware, then he will probably – and deservedly – make better inferences than I will. But an alert mind can recognize, from the failure of its inferences, that additional information was needed, and from the nature of the failure will have a clue as to where to seek it. You seem to think it would be a calamity to leave out a pertinent piece of prior information "lurking out there"; why don't you see the Bayesian formalism as a golden opportunity to *learn* what is pertinent?"

REPLIES TO COMMENTS ON PME

Having spent some thirty years in the development and use of PME methods in physical problems where the realities of the situation could not be ignored, I know very well how Hannibal felt on beholding a philosopher, who had never seen a battle line, discoursing on warfare. The urge to react as Hannibal did is overpowering.

Hannibal's critic would, of course, concentrate on problems that existed only in his imagination, and would be quite unaware of the real circumstances that determine the actual course of a battle. Likewise, some of my critics become mired in worries about such things as the exact meaning of the constraints, but seem unaware of the real difficulties that we encounter in applying PME to new problems.

The main difficulty is, almost always, how to choose the "hypothesis space" on which we define our entropies. I have been held up for years

by unsure judgment in defining an hypothesis space, but have never yet seen a problem in which there was any difficulty in deciding which constraints should be applied.

In statistical mechanics, the hypothesis space problem was solved long ago. Extending Liouville's theorem to quantum mechanics, the linearly independent "global" quantum states of a system define, according to all present knowledge, the proper space on which PME leads, unerringly, to correct predictions. But what takes their place in a problem of econometrics? In trying to apply PME in a new area we are sometimes in a situation corresponding to (if one can imagine it) statistical mechanics before the discovery of Liouville's theorem.

In maximum-entropy image reconstruction, the most obvious hypothesis space has led to important advances; yet as a careful reading of John Skilling's comments may suggest, deeper problems of defining this space are going to arise as we search for future refinements. For example, should we reduce it to individual photons (I think not); or define the smallest cell by our measurement error (perhaps); or should we go into entirely different coordinates that express some correlation between nearby picture elements (probably). In trying to answer these questions, we should look eagerly for evidence of any systematic failure of PME reconstructions based on the present hypothesis space, which would give us a clue to a better one.

In maximum-entropy spectrum analysis only one explicit solution, that of Burg (1967) is available thus far, and although again very important advances have been made with it, the problem of the hypothesis space has hardly been faced as yet. There is no reason to think that the proper space for a geophysical time series must also be right for an econometric one or an ecological one.

These remarks may indicate how irrelevant to actual scientific practice are mathematical/philosophical hangups over issues that do not refer to any specific real problem. In Francis Perry's comments there is a good recognition of this, and some deep thought about it. More thinking along those lines, in a variety of real situations, is needed.

As an interpretation of PME on a space S, Skyrms points out that one can imagine S enlarged to S^n on which Bayes's theorem can be appplied, leading asymptotically to the same mathematical solution. Of course one can do this; indeed, that is just what Darwin and Fowler did in the 1920's, and what Schrodinger (1948) and Eyring (1965) adopted

as the basic principle of statistical mechanics. What this shows is only that two problems can be conceptually different, but so similar mathematically that the same numerical algorithm can be used for both.

* * *

Immediately after writing this last sentence, it struck me that perhaps I am misjudging Skyrms' intentions. It may be that Skyrms was not offering his work as a new contribution to the field, but wanted only to translate something already known into his own language. In that case we are concerned only with the suitability of that language as a description of the real problems where we use PME.

Now in most real problems there is no "random experiment" and no "random variables"; and the extension space S^n is purely a figment of our imagination. Surely, it is inelegant and unnecessary to drag in all these extraneous notions in order to "justify" the PME procedure.

We started with a clear, simple problem in which a clear, simple desideratum (honesty) points to a solution that proves to be feasible to calculate and useful in practice. It seems to me that the important advance and a major virtue of PME is that it gives us this in such a direct way that *avoids* all the clutter that was invoked in the past.

On the other hand, there are some problems in which the space S^n is actually present, in which case application of Bayes' theorem is clearly the fundamentally correct procedure. But to do this exactly for finite n proves to be very tedious mathematically, and in practice one quickly discovers that if n is reasonably large there is an excellent approximation. It is, of course, just that PME solution!

After going through the Bayesian analysis and extracting the asymptotic solution. Darwin-Fowler, Schrodinger, Eyring, and everybody else since the time of Gibbs has reverted, for their actual useful calculations, to the PME algorithm.

The mathematical relations noted by Skyrms are, therefore, quite correct and well-known; but I should draw from them a very different conclusion from his. In the real world, even when that space S^n actually exists, it is almost always the direct application of PME that gives us the pragmatically useful results.

I have already deplored the use of the terms "prior" and "posterior" to describe two PME solutions with different constraints. Skyrms certainly confuses his readers by this, and perhaps also himself.

It is hard to imagine a more unfortunate, ill-advised remark than his parting shot: "Is it good methodology for the blind man to assume that the road is smooth and wide because he cannot see the ruts or the ditch?" This utter falsification of what PME is doing casts a cloud of doubt over everything else in Skyrms' paper. The success of maximum-entropy image reconstruction in *bringing out* detail that other methods have failed to see is one of its most important achievements. In a few years, devices based on PME should be available just to help blind men to perceive those ruts.

GOSSIP OR MEDDLING?

Of course, if philosophers wish to discuss the rationale of science among themselves, in their own journals, without pretending that they are making new contributions to science, they have every right to do so. We physicists also gossip among ourselves about work in other fields – current developments in biology, for example – expressing all kinds of opinions, without thereby pretending that we are making new contributions to biology.

But a physicist, not well informed about the whole general status of the field, would not try to meddle in biology by injecting his own half-formed ideas into the biological journals; for he would almost surely be repeating a line of thought that professional biologists have long since thought of and disposed of; and would only make himself ridiculous in their eyes.

And nothing could be more ridiculous than for a physicist to tell biologists how they ought to mend their ways by attacking particular remarks made by Crick and Watson in 1951, while ignoring everything that Crick proceeded to do, from that beginning, in the 1960's and 1970's.

By the same token, when a philosopher takes it upon himself to move into the scientific journals with criticisms clearly intended to influence scientific practice, then I think he has an obligation to get his technical facts and documentation right, and to inform himself about current activity in the field; otherwise he will at best only make a clown of himself, and at worst do serious damage.

This brings us, obviously, to the matter of Shimony. I am not a participant, but, like other readers, only a bewildered onlooker, in the spectacle of his epic struggles with himself. He seems to have made it

his lifelong career to misconstrue everything I wrote many years ago, and then compose long, pedantic commentaries, full of technical errors and misstatements of documentable fact, showing no awareness of anything done in this field since then – and which, to cap it all off, attack not my statements, but only his own misunderstandings of them. The conflict is not between Shimony and me, but between Shimony and the English language.

I want to defend both myself and Shannon against Shimony's misleading accounts of our work. As anyone can verify, the argument that Shimony attributes to Shannon (1948) leading to Shimony's Eq. (1), omits the appeal to consistency that gives Shannon's argument its force; indeed, if Shannon's derivation had no more substance than the one Shimony reports, the name "Claude Shannon" and the term "Information Theory" would be quite unknown today.

Then he turns to my work. I deny that I have ever defined PME by the statement that Shimony attributes to me. The phrase "The p_i have those values . . . " is anathema, conveying the opposite of my meaning. Then he too commits that error of terminology and notation; maximum-entropy probabilities are not only called "posterior probabilities" but even (Eq. 4) denoted explicitly by the symbol for a conditional probability! In the works of others, this only causes confusion; but for Shimony this sets off a chain reaction of further errors.

The notation which confuses two quite different things in (4) deludes him into failing to see the distinction between the Lagrange multiplier β of a maximum-entropy problem and an estimated parameter of a sampling distribution. This in turn leads him to suppose that he and Friedman have discovered an "anomaly", in the fact that PME leads to the value $\beta = 0$ when there are no constraints other than normalization.

Although the actual result (2) has been a standard part of probability theory for 250 years, his attempts to interpret it in terms of a probability distribution for a Lagrange multiplier leads him to a quite new and startling conclusion. We are not all in hell, although we need not question Shimony's account of his own predicament.

Errors in this argument have now been pointed out five times, by Tribus and Motroni (1972), Hobson (1972), Gage and Hestenes (1973), Jaynes (1978), and Cyranski (1979); yet he persists in publishing that same argument over and over again. So we can hardly hope to enlighten him with anything written here; but in the following remarks addressed to others we point out a few elementary technical facts that

will help to avoid the pitfall in which Shimony seems permanently trapped.

In the maximum entropy problem, the quantity β had no previous existence; it is a Lagrange multiplier that is created only in the process of entropy maximization. But it appears only for mathematical convenience; the problem could be solved also by direct algebraic reduction without ever introducing it. β is not "estimated", but *defined*, by the PME formalism. That it is defined exactly and not approximately, far from being cause for complaint, merely indicates that our maximization problem was mathematically well posed. There would be cause to complain were it otherwise.

It does not make sense, therefore, to speak of having prior knowledge of β, much less of honestly representing that knowledge. A Lagrange multiplier does not have a probability distribution; it is no different, in principle, from a normalization constant that also appears in a probability distribution. That too is not estimated, but defined; and indeed, to infinite accuracy. Would Shimony complain that this too violates the honesty maxim, and demand that henceforth we use inaccurate normalization?

The question is not at all facetious; for the λ_0 in the PME formalism is the Lagrange multiplier that is chosen to fit the normalization constraint, so perhaps he has already done this.

Of course, PME is different from Bayes' theorem, because it addresses a different problem, with a different kind of information, and for a different purpose. A MAXENT distribution is not a "posterior distribution" and we are not making inferences about any parameters in a previously defined sampling distribution. My attempts to point this out (Jaynes, 1978) were not comprehended because of Shimony's seeming inability to read a simple English sentence.

Shimony quotes two statements of mine, which he reports as claiming that a proposition used to provide a constraint in PME cannot be used as a conditioning statement in Bayes' theorem. But I made no such claim, as Shimony might see for himself if he would only read the statements that he quotes. If someone points out a rock that is white but not round, and another that is round but not white, he has not thereby denied that it is possible for a rock to be both white and round; he has merely noted that the examples before him do not have that property.

Indeed, it would hardly be a feat to produce such a double-action proposition, if one had one of type d and one of type D. Their

conjunction *Dd* would seem to do the job, would it not?

Since most of the rest of his discussion is a quixotic attempt to tilt with a claim that was never made, it does not seem worth examining here.

Not satisfied with confusing PME with Bayes' theorem and with parameter estimation, Shimony has also confused it with the Carnap inductive methods and complained that it is not equivalent to any with finite lambda. But in the real world, this difference is one of the main merits of PME. No scientist would use a Carnap method to predict the future of a geophysical or economic time series from its past, because that method presupposes that correlations persist undiminished for all time. Maximum entropy spectrum analysis automatically leads us to realistic predictions in these problems, because it introduces only those correlations (autoregressive coefficients) for which there is evidence in the data; this is just that "honesty maxim" at work (Jaynes, 1982).

Shimony has now published the incredible statement that, because of his lack of understanding, the use of PME "ought to be curtailed", which sounds more like a threat than the observation of a scholar. Those of us who are engaged in constructive activities, in which PME has proved to be a useful tool, will continue to use it with or without Shimony's permission.

CONCLUSION

Of course, the rationale of PME is so different from what has been taught in "orthodox" statistics courses for fifty years, that it causes conceptual hangups for many with conventional training. But beginning students have no difficulty with it, for it is just a mathematical model of the natural, common sense way in which anybody does conduct his inferences in problems of everyday life.

The difficulties that seem so prominent in the literature today are, therefore, only transient phenomena that will disappear automatically in time. Indeed, this revolution in our attitude toward inference is already an accomplished fact among those concerned with a few specialized applications; with a little familarity in its use its advantages are apparent and it no longer seems strange. It is the idea that inference was once thought to be tied to frequencies in random experiments, that will seem strange to future generations.

REFERENCES

Boltzmann, L.: 1877, *Wiener Berichte* **76**, 373.
Cox, R. T.: 1946, *Am. Jour. Phys.* **17**, 1.
Cox, R. T.: 1961, *The Algebra of Probable Inference*, Johns Hopkins University Press, Baltimore, Reviewed by E. T. Jaynes, Am. Jour. Phys. **31**, 66, 1963.
Cyranski, J.: 1979, *Information and Control* **33**, 275–304.
Darwin, G. C. and Fowler, R. H.: 1928, see account in R. H. Fowler (1936), *Statistical Mechanics*, Cambridge University Press.
Eyring, H.: 1964, *Statistical Mechanics and Dynamics*, J. Wiley & Sons, Inc., New York.
Gage, D. and Hestenes, D.: 1973, *J. Stat. Phys.* **7**, 89.
Gibbs, J. W.: 1902, *Elementary Principles in Statistical Mechanics*, reprinted (1961) by Dover Publishing Co., New York.
Hobson, A.: 1972, *J. Stat. Phys.* **6**, 189.
*Jaynes, E. T.: 1971, 'The Well-Posed Problem', in V. P. Godambe and D. A. Sprott, (eds.), *Foundations of Statistical Inference*, Holt, Rinehart & Winston of Canada, Toronto.
*Jaynes, E. T.: 1978, 'Where Do We Stand on Maximum Entropy?', in R. D. Levine and M. Tribus (eds.), *The Maximum Entropy Formalism*, M.I.T. Press, Cambridge MA.
*Jaynes, E. T.: 1980, "Marginalization and Prior Probabilities", in A. Zellner (ed.), *Bayesian Analysis in Econometrics and Statistics*, North-Holland Publishing Co.
Jaynes, E. T.: 1982, 'On the Rationale of Maximum-Entropy Methods', *Proc. IEEE*, **70**, 939–952.
Jeffreys, H.: 1939, 1948, *Theory of Probability*, Oxford University Press, 1st and 2nd editions.
Savage, L. J.: 1954, *The Foundations of Statistical Inference*, J. Wiley & Sons, Inc., New York.
Schrodinger, E.: 1948, *Statistical Thermodynamics*, Cambridge University Press.
Tribus, M. and Motroni, H.: 1972, *J. Stat. Phys.* **4**, 227.

* These articles are reprinted in E. T. Jaynes (1983), *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkrantz, Editor, D. Reidel Publishing Co. Dordrecht-Holland.

Dept. of Physics
Washington University
St. Louis, Missouri 63130
U.S.A.